J Formos Med Assoc 2010;109(3):209–218

Contents lists available at ScienceDirect

Journal of the Formosan Medical Association

Journal homepage: http://www.jfma-online.com

Original Article

Selecting a Cutoff Point for a Developmental Screening Test Based on Overall Diagnostic Indices and Total Expected Utilities of Professional Preferences

*Hua-Fang Liao,[1,2] Ling-Yee Cheng,[3] Wu-Shiun Hsieh,[4] Ming-Chin Yang[5]**

Background/Purpose: A cutoff point in a test with sounded validity and professional preferences can help to make an accurate clinical decision. This study aimed to determine a cutoff point between two strategies for a developmental screening checklist (referred to as Taipei II). Cutoff point A was set as one or more item failed and cutoff point B was set as two or more items failed or one or more marked item failed.

Methods: This study was based on the total expected utilities of professional preferences and overall diagnostic indices. A self-administered questionnaire was developed to collect the estimated utility from professionals involved in early childhood interventions ($n = 81$) regarding four screening outcomes (probabilities of true positive, false positive, true negative, or false negative) and costs. The total expected utilities were calculated from the probabilities of four screening outcomes and utility values.

Results: The diagnostic odds ratio was higher for strategy B (695 and 209, respectively) than that of strategy A (184 and 150, respectively) when using the Taipei II on children under 3 years of age and age 3 and over. Strategy B also had a higher median total expected utilities score than strategy A (0.78 *vs.* 0.72 for children < 3 and 0.76 *vs.* 0.67 for children ≥ 3).

Conclusion: If only one cutoff point can be chosen, the authors suggest that clinicians should choose cutoff point B when using the Taipei II for screening. However, two cutoff points of Taipei II, a combination of strategy A and B, can also be used clinically.

Key Words: child, developmental delay disorders, screening, Utility theory

An increasing emphasis has occurred over the past 30 years on the early identification and minimization of developmental delays in children.[1] If clinicians rely solely on clinical judgment, less than one half of the cases of mild mental retardation or emotional/behavioral disorders are identified.[2] Therefore, a cutoff point in a developmental screening test with sounded reliability and

[1]School and Graduate Institute of Physical Therapy, [4]Departments of Pediatrics, College of Medicine, and [5]Graduate Institute of Health Care Organization Administration, College of Public Health, National Taiwan University; [2]Physical Therapy Center, Department of Rehabilitation Medicine, National Taiwan University Hospital, and [3]Department of Physical Medicine and Rehabilitation, Taipei Veterans General Hospital, Taipei, Taiwan.

Received: January 12, 2009
Revised: June 18, 2009
Accepted: July 1, 2009

***Correspondence to:** Dr Ming-Chin Yang, Room 637, 17 Xuzhou Road, Taipei, Taiwan.
E-mail: mcyang637@ntu.edu.tw

validity is helpful for the accurate identification of developmentally-delayed children.[3,4] There are tradeoffs, however, between sensitivity and specificity in different cutoff points.[5] Variations in diagnostic criteria and target population affect the sensitivity and specificity of a developmental test.[6] Thus, the choice of cutoff point not only depends on psychometric properties, but also subjective value judgments about how to weigh the adverse effects, such as the cost of incorrect diagnosis [false positive (FP) or false negative (FN)] versus the beneficial effects of correct diagnosis [true positive (TP) or true negative (TN)].[5,7] However, most of the consequences or costs after the application of developmental screening tests, such as life-years gained or disabled-years saved, that are usually used in the classical economic evaluation forms have not been defined.[8] Fortunately, subjective utilities may be estimated by relatively objective cost-accounting procedures in the decision-making strategy.[9] Therefore, the harm or benefit of the screening test can be estimated by the preferences and value system of a specific group or general population. A best cutoff point of a screening test can then be chosen from the total expected utilities (TEU) of different cutoff strategies.[4]

For the application of screening tests in a population, it is suggested that a policy advisory committee should determine the national policy for screening implementation, such as the cutoff point of each screening test.[10] Members of that committee are usually early intervention-related professionals such as clinicians, public health providers, therapists, educators and social workers. However, no previous study has shown the expenses and outcomes (TP, TN, FP, FN) of developmental screening tests in various professions. Thus the authors designed a questionnaire to collect the utility estimations of screening outcomes from early intervention professionals and to examine the reliability and the influencing factors of these utility estimations.

The Taipei City Developmental Checklist for Preschoolers, 2nd version (Taipei II), revised in 2005,[11] is a concise screening instrument aimed at identifying children who should receive further assessment because of potential risk of developmental delays or disabilities. The Taipei II provides 13 checklists for 13 age groups: 4, 6, 9, 12, 15, 18, 24, 30, 36, 42, 48, 60, and 72 months. Each checklist has 11 to 13 behavior- or skill-related items. Gross and fine motor control, cognition, language/communication, and emotional/social areas are easily observed or elicited by the child's caregiver. The internal consistency coefficients (α) of the 13 checklists of the Taipei II are 0.72 – 0.87. Taipei II has been applied widely in Taiwan in recent years.[12]

Two cutoff points (A and B) have been suggested for the Taipei II with the data-based probabilities of four screening outcomes. Cutoff point A was set as more than one item failed while cutoff point B was set as more than two items failed or more than one marked item failed.[12] Validity studies of the Taipei II from a sample of 3792 children aged 4–72 months in a community setting ($n=3146$) or medical care institutes ($n=646$) showed that the sensitivity ranged 0.85–1.00 and specificity 0.82–1.00 for cutoff point A. For cutoff point B, the sensitivity ranged 0.75–1.00 and specificity 0.72–1.00. (Dr L.Y. Cheng, written communication, January 1, 2007) However, the better cutoff point is yet to be determined.

Thus the purposes of this study were to investigate the reliability and the influencing factors on the expected utilities of the screening outcomes. The secondary aim was to compare the TEU and the overall diagnostic indices of the two cutoff points for the Taipei II and to choose a better cutoff strategy based on the psychometric properties of the Taipei II and professional preference.

Materials and Methods

This study used a self-administered utility questionnaire to survey various professionals to obtain the expected utilities of four screening outcomes (TP, TN, FP, and FN) and the expenses of conducting a screening test. The TEUs of two strategies of the Taipei II were then calculated

from the probabilities of four screening outcomes and utility values. Secondary studies were performed for overall diagnostic indices of the two strategies of the Taipei II based on data from a community and consecutive sample.

Study tool and participants

The principle of maximization of the decision theory, to maximize the TEU of all possible outcomes, namely TP, TN, FP, and FN, was used for the design of the questionnaire.[9] To maximize an averaged TEU, utility values must be expressed in terms of comparable units.[9] Previous studies showed that the visual analog scale (VAS) was more sensitive to small changes than ordinal scales,[13] and to avoid the floor effect of scaling, the graphic rating VAS was the best choice.[14] The graphic rating VAS is a graduated verbal descriptive scale, with 10 cm gradations, followed by descriptive terms along the line. In this study, the five descriptive terms along the line from −50 mm to +50 mm were: very bad, bad, not good or bad, good, very good (see Appendix).

The questionnaire was sent to 19 professionals twice at an interval of 1 month to examine the test–retest reliability. There were 14 medical professionals (pediatricians, clinical psychologists, and therapists) and five non-medical professionals (social workers, teachers). They consisted of 17 females, with an average age of 35.8 ± 10.4 years and 6 years experience in early childhood intervention.

After the reliability test, the self-administered questionnaire was sent to 84 professionals in Northern Taiwan to estimate their utilities. Eighty-one effective questionnaires were obtained. Among the respondents, 51 were medical professionals (21 pediatricians, 4 psychologists, 14 therapists, 1 nurse, and 11 public health professionals) and 30 were non-medical professionals (16 social workers, 11 teachers, and 3 health administrators; Table 1).

Study of the screening outcomes of the Taipei II

Existing empirical data on the Taipei II was used to calculate the probabilities of the four screening outcomes. The data came from a consecutive sample of 3146 children aged 4 – 72 months in a community setting in Northern Taipei, including well baby clinics of two hospitals, five local primary care units, and 30 preschools or kindergartens. The authors chose the community sample to calculate the screening outcomes of the Taipei II because it reflected the real screening procedure in Taiwan. Most of the Taipei II checklists were completed by clinical psychologists after interviewing children's caregivers and testing/observing children directly. Some checklists were completed by children's caregivers and checked by clinical psychologists. The external criterion

Table 1. Demographic data and estimated utilities of the professionals for utility estimations*

	Total ($n = 81$)	Medical ($n = 51$)	Non-medical ($n = 30$)	p
Age (yr)	34.7 ± 9.1	35.6 ± 10.0	33.3 ± 7.3	590.5 (0.44)†
Sex, female	61 (75.3)	33 (64.7)	28 (93.3)	6.55 (0.01)‡,§
With screening experience	68 (84.0)	40 (78.4)	28 (93)	2.61 (0.11)‡
Pediatric experience (yr)	4.5 (1.5–8.0)	4.0 (1.0–10.0)	4.5 (2.5–7.0)	678.5 (0.97)†
Estimated utilities				
True positive	0.8 (0.6–0.98)	0.8 (0.6–1.0)	0.75 (0.6–0.8)	597.5 (0.09)†
True negative	0.6 (0.4–0.95)	0.6 (0.4–0.8)	0.7 (0.4–1.0)	723.0 (0.68)†
False positive	−0.6 (−0.8 to −0.4)	−0.6 (−0.8 to −0.4)	−0.4 (−0.65 to −0.08)	544.5 (0.03)†,§
False negative	−0.8 (−1.0 to −0.6)	−0.8 (−1.0 to −0.6)	−0.7 (−0.93 to −0.55)	552.0 (0.03)†,§
Expenses	0.2 (0–0.45)	0.2 (−0.2–0.4)	0.25 (0.18–0.6)	628.0 (0.18)†

**Data presented as mean ± standard deviation, n (%), or median (interquartile range); †by Mann-Whitney test, medical versus non-medical professionals; ‡by Pearson χ^2, medical versus non-medical professionals; §$p < 0.05$.*

for the delay was suspected delay or delay judged by a multidisciplinary team including medical professionals and teachers.

Statistical analysis

Data were analyzed by SPSS version 11.0 (SPSS Inc., Chicago, IL, USA). The normality of variables was examined first by the Shapiro-Wilks test. For the 1st test, the *Ws*[18] of the five utility estimations, except the expenses utility, ranged from 0.55 to 0.95 ($p = 0.000 - 0.458$) and was, therefore, against the normal distribution assumption. The Spearman correlation coefficient was then used to test the test–retest reliability (stability) of the utility estimation. For analysis of influential variables on utilities estimation, normality of utilities estimation of the whole group was examined first by Kolmogorov-Smirnov test. All utility variables were against the normal distribution assumption (K-S statistic = 0.15 – 0.22, $p < 0.001$). Therefore, non-parametric statistic tests were used to explore influential factors on utility values.

The overall diagnostic indices used in this study were the Youden index (YI) and the diagnostic odds ratio (DOR). The YI is calculated using the formula:

$$\text{YI} = \text{sensitivity}\ (\%) + \text{specificity}\ (\%) - 100,$$

and is independent of prevalence. The larger the YI, the better the validity of the test. YI equal to 0 indicates a useless test.[15] The DOR is the ratio of the odds of a positive result with disease relative to the odds of a positive result with non-disease.[16] The DOR can be calculated by the following formula:

$$\text{DOR} = (\text{TP}/\text{FN})/(\text{FP}/\text{TN}) = (\text{positive likelihood ratio})/(\text{negative likelihood ratio}).$$

The value of the DOR ranges from 0 to infinity. A higher DOR value indicates good separation between a positive and negative test. A DOR value < 1 reveals improper test interpretation.[16] The minimum acceptable value of DOR is 50, and a value > 500 is considered very good.[16] For clinical application, the positive likelihood ratio (LR+) and negative likelihood ratio (LR–) were also calculated in this study, as was the 95% confidence interval (CI) of DOR.[16] The LR indicates the likelihood of a given test result in a patient with the target disorder compared with the likelihood of the same result in a patient without that disorder.[16] The formula of LR for a positive test result is:

$$\text{LR}+ = \text{sensitivity}/(1 - \text{specificity});$$

and that of the LR for a negative test result is:

$$\text{LR}- = (1 - \text{sensitivity})/\text{specificity}$$

LRs > 10 or < 0.1 generate large and often conclusive changes from pretest to posttest probability. LRs of 5–10 and 0.1–0.2 generate moderate shifts in pretest to posttest probability. LRs of 2–5 and 0.5–0.2 generate small (but sometimes important) changes in probability. LRs of 1–2 and 0.5–1 alter probability to a small (and rarely important) degree.[17]

The TEU of two cutoff points of the Taipei II were calculated by using the averaging out method,[5] multiplying the probabilities by the outcome utility for each of the events as in the following formula:[4]

$$\text{TEU} = P_{tp} \times U_{tp} + P_{fp} \times U_{fp} + P_{tn} \times U_{tn} + P_{fn} \times U_{fn} + U_{exp}$$

P_{tp}, P_{fp}, P_{tn}, and P_{fn} are probabilities of TP, FP, TN and FN, respectively. U_{tp}, U_{fp}, U_{tn}, U_{fn}, and U_{exp} are utilities of TP, FP, TN, FN and expenses of the screening application, respectively. Based on the result of each questionnaire, we obtained the five VAS values (mm) of four screening outcomes (TP, TN, FP, FN) including VAS_{tp}, VAS_{tn}, VAS_{fp}, VAS_{fn}, and the expense (VAS_{exp}). The U_{tp} was calculated according to the following formula:

$$U_{tp} = VAS_{tp}/50.$$

The "50" is the absolute value in the VAS. The U_{tn}, U_{fp}, U_{fn}, and U_{exp} were calculated in a similar fashion. For sensitivity analysis, we used two

methods to estimate possible ranges of the TEUs. We first used the maximum and minimum value to calculate possible ranges of the five VAS values. Alternatively, because the responses were not normally distributed, we used bootstrapping to draw 1000 bootstrap samples and generate the upper and lower limits of the 95% CI of the five VAS values. We then calculated the minimum and maximum levels of the TEUs of the two cutoff points.[8] The prevalence of developmental delay will change the magnitudes of P_{tp}, P_{fp}, P_{tn}, P_{fn}, and in turn, the TEU.[15] An estimated prevalence of developmental delay among preschool children has been previously shown to be 4–9%.[18] The prevalence rate of developmental delay and suspect delay range were then estimated to be 4–18%. The minimal and maximum prevalence rates were also used for sensitivity analysis.

Results

Test–retest reliability of the utilities estimation

The U_{tp}, U_{fp}, U_{fn}, and U_{exp} of the two repeated tests were significantly correlated ($r_s = 0.47 - 0.66$, $p = 0.002 - 0.043$). This significance was not seen, however, for the borderline reliability of the U_{tn} ($r_s = 0.45$, $p = 0.051$; Table 2). On further examination of the U_{tn} at the 1st and 2nd test, only two professionals ticked the 0 value (neither good nor bad) at the first test, the other 17 all ticked positive values at two repeated tests, and the test–retest agreement was 89.5%. Therefore, we concluded that the stabilities of the utilities estimation were acceptable.

Factors influencing the utilities estimation

As shown in Table 1, medical professionals viewed FP and FN as more harmful than non-medical professionals did. Also females U_{fn} (median = –0.8, IQR = –1.0 to –0.6) were significantly higher than males (median = –1.0, IQR = –1.0 to –0.8). Female professionals viewed FN as less harmful than male professionals did.

Probabilities and expected utilities of the expenses and outcomes of the screening test

The P_{tp}, P_{fp}, P_{tn} and P_{fn} as well as sensitivity, specificity, LR+ and LR– of two cutoff points for children of two age groups (older than 3 years and younger than 3 years) were calculated from the databank of a previous study, (Dr L.Y. Cheng, written communication, January 1, 2007) and shown in Table 3. The median utilities and inter-quartiles of the four screening outcomes and expenses are shown in Table 1. The median U_{tp}, U_{tn}, U_{fp}, U_{fn}, and U_{exp} was +0.8, +0.6, –0.6, –0.1, and +0.2, respectively (Figure). The professionals valued TP as the highest positive preference and FN as the highest adverse effect. The values of sensitivity and specificity were all above 0.8. The LR+ of strategy B was above 10 and LR– of strategy A was less than 0.1.

Diagnostic indices and total expected utilities of two cutoff strategies of the Taipei II

In both age groups the YIs were all above 0 and ranged from 79 to 83, DORs were above 50 and ranged from 141 to 541 (95% CI: 26–3781). Median TEUs above 0 ranged from 0.67 to 0.78 (Table 4). For children aged less than 3 years, the DOR of strategy B was above 500, excellent for

Table 2. Test–retest reliabilities of the expected utilities of four screening outcomes and screening expenses ($n = 19$)*

	1st test	2nd test	r_s†	p
True positive	0.8 (0.8–1.0)	0.8 (0.8–1.0)	0.47	0.043
True negative	0.8 (0.6–1.0)	0.8 (0.6–0.8)	0.45	0.051
False positive	–0.6 (–0.6 to –0.2)	–0.6 (–0.8 to –0.4)	0.65	0.003
False negative	–0.8 (–1.0 to –0.6)	–0.8 (–1.0 to –0.8)	0.51	0.027
Expenses	0.2 (0.0–0.6)	0.4 (0.2–0.6)	0.66	0.002

**Data presented as median (interquartile range); †Spearman correlation coefficient.*

Table 3. The validity indices of the Taipei II in the community sample

Age (yr)	Cutoff*	Probabilities†				Sensitivity	Specificity	Positive LR	Negative LR
		TP	TN	FP	FN				
<3	A	0.129	0.767	0.097	0.007	0.95	0.89	8	0.06
(*n*=1104)	B	0.109	0.858	0.006	0.027	0.80	0.99	109	0.20
≥3	A	0.162	0.702	0.131	0.005	0.97	0.84	6	0.04
(*n*=2042)	B	0.136	0.815	0.017	0.032	0.81	0.98	39	0.19

**Strategy A: number of failure items ≥1, strategy B: no. of failure items ≥2 or failure star items ≥1; †in a written communication with Dr L.Y. Cheng, January 1, 2007, a community and consecutive sample, criteria reference is developmental delay or suspect delay judged by clinicians. TP=True positive; TN=true negative; FP=false positive; FN=false negative; LR=likelihood ratio.*

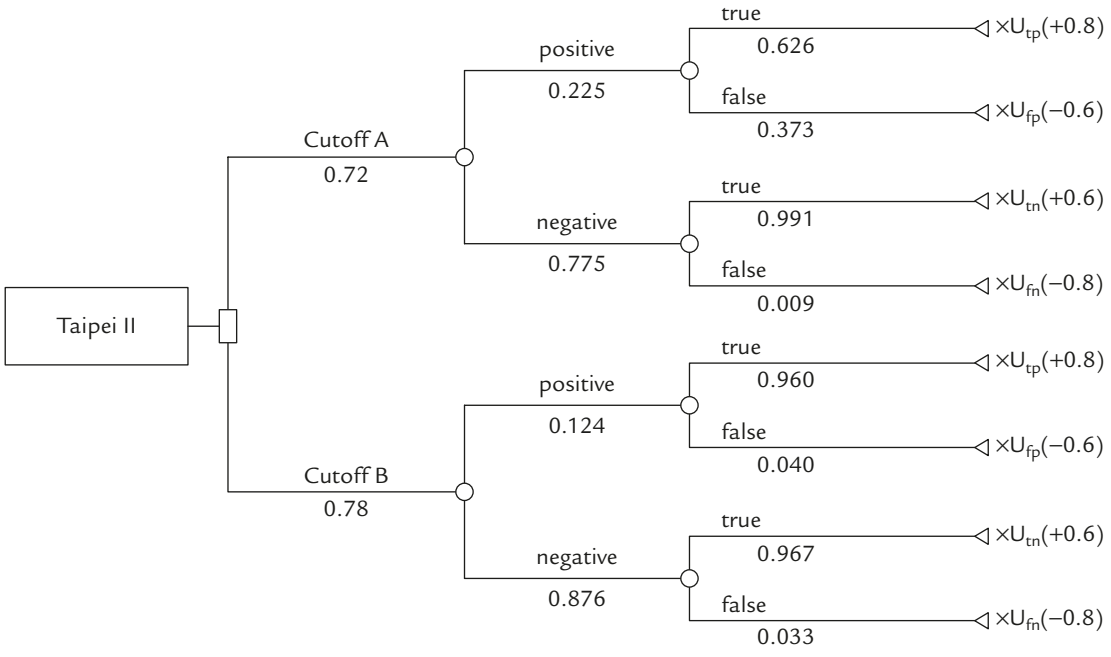


Figure. Decision tree comparing two cutoff strategies of the Taipei II by calculating total expected utilities using children less than 3 years old as a example.

differentiating positive and negative test results. For both age groups, the cutoff strategy B showed higher DOR and TEU than strategy A. However, strategy A had higher YI than strategy B. In terms of sensitivity analysis, the TEUs ranged from 0.57 to 0.97 for strategy B and from 0.49 to 0.91 for strategy A. This suggests that the TEUs of strategy B were slightly higher than that of strategy A, although a wide range of overlap existed.

Discussion

The DORs and the TEU values determined in this study revealed that cutoff strategy B tended to be better than strategy A. Although a wide range of overlap existed between the 95% CI of DOR, the value for strategy A covered the values less than 50, and that for strategy B was larger than 50 in both age groups. From the YI values, cutoff point A was better than point B as using YI to choose the optimal cutoff strategy is usually under the assumption that equal weight is given to TP, TN, FP and FN.[19] However, in this study, the weight for the four screening outcomes was different. Therefore, if only one cutoff point were to be chosen, either from the overall diagnostic indices or the TEUs, cutoff point B is better than point A. However, for clinical decisions, the multilevel LRs of a test or a screening strategy are more powerful and

Table 4. The diagnostic indices and total expected utility of the Taipei II in two age groups

Age (yr)	Cutoff*	YI (%)	DOR (95% CI)	Total expected utilities				
				Minimum PR	Maximum PR	Bootstrapping		
						Median	Minimum	Maximum
<3	A	83	141 (26–772)	0.71	0.68	0.70	0.53	0.91
	B	79	541 (77–3781)	0.77	0.79	0.78	0.58	0.97
≥3	A	81	161 (39–665)	0.67	0.63	0.67	0.49	0.87
	B	79	203 (75–546)	0.76	0.77	0.76	0.57	0.95

**Strategy A: no. of failure items ≥ 1, strategy B: no. of failure items ≥ 2 or failure star items ≥ 1. YI = Youden index; CI = confidence interval; DOR = diagnostic odds ratio; PR = prevalence rate.*

useful than one single cutoff point.[15] The high magnitude of LR+ of strategy B (>35) as well as the low value of LR– of strategy A (<0.1) could provide conclusive changes from pretest to posttest probability in clinical decision making for either positive test results or negative test results. Therefore, two cutoff points of the Taipei II could be used clinically. For children who fail ≥ two items or one marked item further diagnosis is recommended. For children who pass all items in the Taipei II, no further assessment is needed. For children who fail one non-marked item, a second screening or closer monitoring is recommended. The multilevel LRs of the Taipei II need further study.

To the authors' knowledge, the present study is the first to use a utility questionnaire to obtain the preferences for outcomes and expenses of screening tests. As mentioned by Wiggins, even within the field of economics, measures of "subjective utility" are frequently preferable to measures of "objective utility".[9] The test–retest reliability study demonstrated that such preference estimates have acceptable stability. Among all the screening outcomes and expenses, the utility value for the TP was the highest and the FN was the lowest. This means that professionals prefer to have a screening test with higher TP and lower FN, even with the slight sacrifice of elevated expenses. From the TP and FN values in Table II, strategy A is better than strategy B. Therefore, if only one or two validity indices are considered, cutoff strategy A is a better choice. However, when using the TEU, all the probabilities and utilities will be merged together to obtain a best choice—in this case strategy B. The prevalence rates of developmental delays will also influence the magnitude of TEU, and not only the utility values. The prevalence rate was 16.7% in the original databank. If we assumed the prevalence rates to be 4%, then for the over-3 age group using cutoff strategy A, the P_{tp} and P_{fn} will decrease to 0.039 and 0.001, respectively, with higher P_{tn} (0.809) and P_{fp} (0.151). For cutoff strategy B, the P_{tp}, P_{tn}, P_{fp}, and P_{fn} were 0.032, 0.940, 0.020, and 0.008, respectively. The TEUs of strategy A and strategy B were 0.63 and 0.77, respectively. The TEU of strategy B was still higher than that of strategy A.

In general, the utility of expenses of administering a test is usually a negative value.[9] In this study, most professionals assigned a positive value for the expense utility. According to the descriptions by professionals, benefits of the high expense of the screening test included: able to train testers to increase the screening reliability and validity and to provide consultation for parents. Therefore, expenses are less of a concern to professionals in Northern Taiwan than the validity of the screening test.

This study found that sex and profession were the only influential factors on the utility estimations. Female professionals estimated FN less harmful effects than their male counterparts. Nonmedical professionals also estimated FN less harmful effects than medical professionals. The reason why sex and profession influence utility values needs further investigation.

Clinicians administering developmental screening tests need to know the predictive abilities of the tests to correctly interpret and communicate the significance of a child's score to parents. In general, sensitivity levels of 70% or more are acceptable,[20] to limit the number of FN.[21] Specificity levels of 70–80% are realistic, although some experts recommend nothing less than 90% as an acceptable level.[20] As shown in Table 3, the Taipei II has a sensitivity ranging from 80% to 97% and a specificity ranging from 84% to 99%. Furthermore, this screening tool only needs on average 10–20 minutes to complete. Given these observations, the Taipei II is an appropriate screening tool for detecting developmental delay in children aged less than 6 years.

Before the formal questionnaire was used for data collection on utility estimation, the authors tried several versions and failed. For uncertain value of preference measures, the standard gamble is usually suggested.[8] However, the standard gamble is only suitable for measuring health status. We could not apply the standard gamble because the outcomes of the developmental screening test are not health status. A paired comparison technique,[22] such as the "Analytic Hierarchy Process"[23] was also used in one of our pilot studies. The results were all positive values of the estimation, and the authors were unable to differentiate good or bad effects of the four screening outcomes. We tried to provide the probabilities of the four outcomes of the Taipei II under two cutoff strategies for professionals and asked them to estimate monetary utilities. However, professionals completely failed to offer the estimations. The authors then used the direct estimation method with the VAS to collect the utilities and found that this estimation method was relatively reliable. Therefore, the authors suggest that the direct estimation method with VAS may be used for different medical procedures in the future.

Limitations

There are two limitations in this study. First, the relevant values to weigh the different outcome probabilities are better achieved surveying the clients, not professionals.[5] However, at this moment parents in Taiwan are not familiar with the terms of probabilities such as TP, TN. According to Hauser-Cram et al, the outcomes of an early intervention program should evaluate from various points of view.[24] Therefore, the authors collected preferences of screening outcomes from various professionals related to early childhood intervention. Further studies for the collection of such information from the societal view point are recommended. Second, the professional sample is not stratified from the total professional population. Therefore, the utility estimations may not represent all professions related to the screening test user. However, this study recruited various professions related to early childhood development and found that only few factors influence the utility values. We also used bootstrapping to generate the 95% CI of five VAS values to estimate the minimum and maximum of the utility estimations. Therefore, the representation problem of the professional sample in this study was resolved in an acceptable manner.

In conclusion, this study found that a utility questionnaire with graphic rating VASs could obtain a reliable preference or utility estimation for screening outcomes in professionals related to childhood early intervention. The professionals preferred to have a screening test with higher TP and lower FN probabilities. If only one cutoff point can be chosen, psychometric properties (DOR) and professional preferences (TEU) show that cutoff strategy B is better than strategy A for clinical application of the Taipei II. However, two cutoff points of Taipei II, combination of strategy A and B, can be used clinically.

Acknowledgments

The authors would like to express gratitude for grant support from the Health Bureau, Department of Health, Executive Yuan of the Taiwan (R.O.C) (DOH95-HP-1205). We also express appreciation to the surveyed professionals for help with data collection.

References

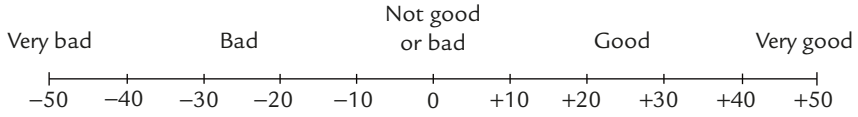
1. Shoemaker OS, Saylor CF, Erickson MT. Concurrent validity of the Minnesota child development inventory with high-risk infants. *J Pediatr Psychol* 1993;18:377–88.
2. Bierman JM, Connor A, Vagge M, et al. Pediatricians' assessments of the intelligence of 2-year-olds and their mental test scores. *Pediatrics* 1964;34:680–90.
3. Brenneman SK. *Assessment and testing of infant and child development*. In: Tecklin, JS. Pediatric Physical Therapy. 3rd ed. USA, Philadelphia: JB Lippincott; 1999:28–70.
4. Anatasi A, Urbina S. *Psychological Testing,* 7th ed. Upper Saddle River, NJ: Prentice Hall; 1997:48–145.
5. Hunink M, Glasziou P, Siegel J, et al. *Decision making in health and medicine: integrating evidence and values*. UK, Cambridge: University Press; 2001:61–213.
6. Baxter P. Normality and abnormality. *Dev Med Child Neurol* 2006;48:867.
7. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2000:79–110.
8. Drummond MF, Sculpher MJ, Torrance GW, et al. *Methods for the economic evaluation of health care programmes,* 3rd ed. New York: Oxford University Press, Inc; 2005: 7–17, 97–209.
9. Wiggins JS. *Personality and prediction: principles of personality assessment*. Reading, Massachusetts: Addison-Wesley Publishing Company;1973:223–74.
10. Muir Gray JA. Testing a screening test;1994. Available at http://www.jr2.ox.ac.uk/bandolier/band5/b5-1.html. [Date accessed: May 1, 2007]
11. Taipei City Developmental Checklist for Preschoolers. Available at http://www.tpscfddc.gov.tw/medicine/check.htm [Date accessed: January 10, 2009]
12. Liao HF, Yang MC, Cheng LY, et al. The cost of the two cut-off strategies of the Taipei City Developmental Screening Checklist for Preschoolers 2nd version. *Formos J Med* 2009;13:9–22 [In Chinese].
13. Keels KD. Pain chart. *Lancet* 1948;2:6–8.
14. Paul-Dauphin A, Guillemin F, Virion JM, et al. Bias and precision in visual analogue scales: a randomized controlled trial. *Am J Epidemiol* 1999;150:1117–27.
15. Straus SE, Richardson WS, Glaszion P, et al. *Evidence-Based Medicine: How to Practice and Teach EBM*, 3rd ed. London: England: Elsevier Churchill Livingstone; 2005:67–99.
16. Glas AS, Lijmer JF, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129–35.
17. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: What are the results and will they help me in caring for my patients? *JAMA* 1994;271:703–7.
18. Liao HF, Wang TM, Yao G, et al. Concurrent validity of Comprehensive Developmental Inventory for Infants and Toddlers with Bayley Scales of Infant Development-II in preterm infants. *J Formos Med Assoc* 2005;104:731–7.
19. Perkins NJ, Schisterman EF. The inconsistency of "optimal "cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* 2006; 163:670–5.
20. Glascoe FP. Parents' concerns about children's development: prescreening technique or screening test? *Pediatrics* 1997;99:522–8.
21. Aylward G. Conceptual issues in developmental screening and assessment. *J Dev Behav Pediatr* 1997;18:340–9.
22. Streiner DL, Norman GR. *Health Measurement Scales, A Practical Guide to their Development and Use*. Oxford, UK: Oxford University Press; 1989:20–38.
23. Saaty TL. *The Analytic Hierarchy Process: Planning, Priority and Resource Allocation*. New York: McGraw-Hill;1980.
24. Hauser-Cram P, Warfield ME, Upshur CC, et al. An expanded view of program evaluation in early childhood intervention. In: Shonkoff JP, Meisels SJ. eds *Handbook of Early Childhood Intervention*, 2nd ed. Cambridge: Cambridge University Press; 2000:487–509.

Appendix. Survey for collecting the utility values of the outcomes and expenses of the developmental screening process

The following questions are about your opinion for the outcomes and expenses of the developmental screening process. There are four possible outcomes for a screening test: true positive, true negative, false positive, and false negative. There is no standard answer for each question. Please answer the questions according to your subjective judgment.

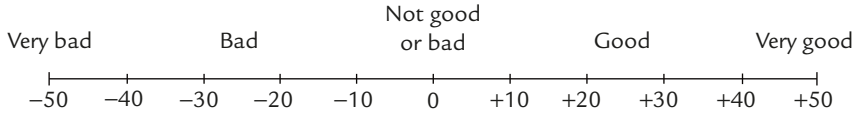
1. If children are truly developmentally delayed and they have been screened and found to be developmentally delayed by a screening test, this would be a true positive. Please describe the benefits or consequences for the children and their families for a screening test with high true positives. _____

   Please mark on the linear scale below the degree of good or bad effects on the children and their families.



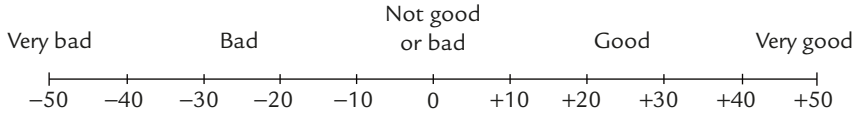
2. If children are truly experiencing typical development (normal children) and they have been screened and found to be typically developed, with no developmental delay, by a screening test, this would indicate a true negative. Please describe the benefits or consequences for the children and their families for a screening test with high true negatives. _____

   Please mark on the linear scale below the degree of good or bad effects on the children and their families.



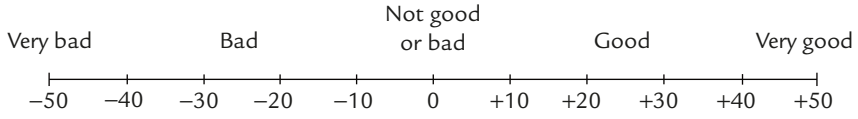
3. If children are truly developmentally delayed, but they have been screened and found to be typically developed by a screening test, this would be a false negative. Please describe the benefits or consequences for the children and their families for a screening test with high false negatives. _____

   Please mark on the linear scale below the degree of good or bad effects for the children and their families.



4. If children are truly typically developed, but they have been screened and found to have a developmental delay by a screening test, this indicates a false positive. Please describe the benefits or consequences for the children and their families for a screening test with high false positives. _____

   Please mark on the linear scale below the degree of good or bad effects for the children and their families.



5. For early detected children with developmental delay, there are financial and time expenses related to administering a screening test, such as test purchasing, tester training, screening test application and explanation of the test results by professionals. Please describe the benefits or consequences for the children and their families for a screening test with high expenses. _____

   Please mark on the linear scale below the degree of good or bad effects for the children and their families.